REMARKS

This amendment responds to the Office Action mailed July 10, 2008, in which the Examiner reopened prosecution. In the Office Action the Examiner:

- rejected claims 12-17, 40, 42-48 and 50-55 under 35 U.S.C. § 103(a) as being unpatentable over Meyerzon et al. (US 6,547,829) in view of Cho et al. ("Finding replicated web collections," Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 355-366, 2000) and further in view of Wang ("Web search services," University of Science and Technology, Hong Kong, issued 2002);
- rejected claims 18-20, 37-39 and 56-58 under 35 U.S.C. §103(a) as being unpatentable over Meyerzon et al. in view of Cho et al. and further in view of Rujan et al. (US 6,976,207) and further in view of Wang; and
- rejected claim 49 under 37 U.S.C. §103(a) as being unpatentable over Meyerzon et al. in view of Cho et al and further in view of Wang and further in view of Lambert et al. (US Pub. No. 2002/0038350).

After entry of this amendment, the pending claims are: claims 12-20, 37-40 and 42-58.

Independent Claims 12, 18, 37, 40, 50, and 56 are patentable over the previously cited references Meyerzon and Cho, as well as the newly cited reference Wang

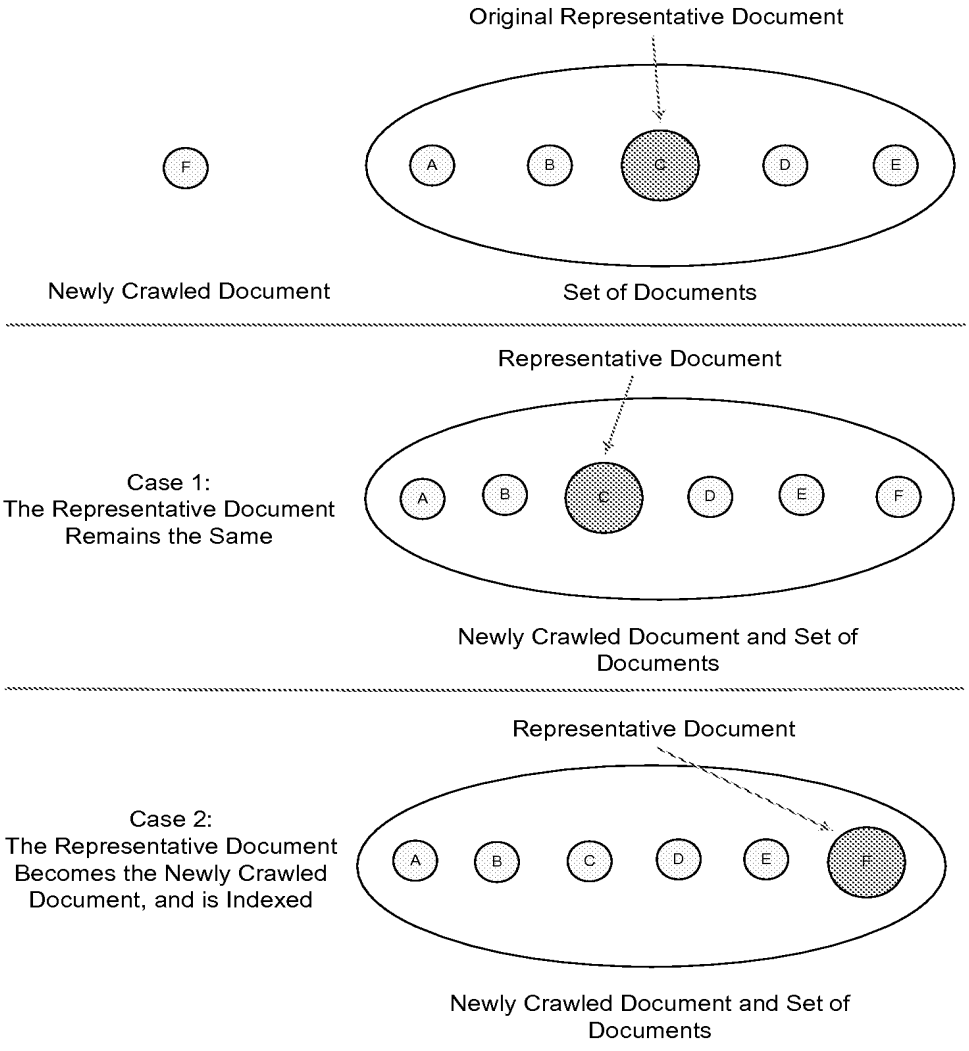
A. **A "representative document" is the one indexed and thus presented to a user.**

Indexing is the operation that makes documents available to a search engine. By indexing a representative document when there are duplicates, the system saves processing resources. In addition, indexing a representative document from each set of duplicates provides a better user experience in response to a query: diverse results are not crowded out by duplicates. Indexing a representative document from each set of duplicates is how the claimed invention achieves its results.

Indexing a representative document is recited in the claims. The claims require "determining a representative document for the newly crawled document and the identified set of documents" and "indexing the representative document when the representative

document is the newly crawled document." This makes sense. The documents in the "identified set of documents" all have the same document content as the newly crawled document. If the newly crawled document becomes the representative, then it needs to be indexed; but if the representative document stays the same, the newly crawled document does not need to be indexed.

The diagram below depicts this process graphically. The top portion shows newly crawled document F, and the set of documents A, B, C, D, and E, all sharing the same content as document F. Here, document C is shown as the original representative document for documents A, B, C, D, and E. Next, document F is added to the set, and a representative selected. The middle portion of the diagram shows the case where document C remains the representative. The bottom portion of the diagram shows the case where document F becomes the representative. In this case document F is indexed.



B. **The claims require that the representative document changes for some documents.**

The claim language "such that at least some of the newly crawled documents are determined to be representative documents and are indexed" conveys the point that the representative document changes for some of the documents. Importantly, the claims address the case where the newly crawled documents are duplicates of documents already known, so selecting the newly crawled document as the representative changes the representative.

For each newly crawled document, the reading operation identifies an original representative document with the same content as the newly crawled document. The original representative document is not the newly crawled document because the original representative document was ascertained from a set of documents already stored in tables.

Therefore, when newly crawled documents are "determined to be representative documents and are indexed," the representative documents have changed.

C. **Meyerzon does not teach a web crawling methodology where the representative document can change.**

Meyerzon addresses the detection of duplicate documents, but responds with a "first copy wins" approach. Meyerzon explains this in the specification at column 9, lines 33-40, with reference to figure 3. When a document is crawled, the crawler determines if the content of the newly crawled document matches the content of a document already in the history table.[1] If the content already exists, then the address (URL) of the newly crawled document is just saved to the history table. If it is not found, then several steps are performed, including step S25, which indexes the new document. Because the first copy of a document is always the one that is indexed, there is no discussion of representative documents, or changing the representative document.

In addition, the Examiner pointed out that Meyerzon does not teach an important limitation in the claims. In the Office Action mailed 08/20/2007, the Examiner stated on page 4 that "Meyerzon does not explicitly indicate:

indexing the representative document when the

[1] Meyerzon uses a "CID," which is defined as a "content identifier." See abstract; column 2, lines 65-67.

representative document is the newly crawled document; and
repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed. (Claim 12)

D. **Cho does not teach a web crawling methodology where the representative document can change.**

Cho teaches detection of duplicate documents, but the first copy found remains the permanent representative. Cho refers to this methodology as a "replica avoiding process." Cho, last paragraph of § 5.1. Specifically, "each crawl identifies new replicated collections that can be avoided in the future." Like Meyerzon, the first copy discovered is the one that is indexed and used. This section of Cho does not teach changing and indexing a representative document.

Cho also presents a revised way to display query results, but does not suggest determining and indexing a new representative. Cho § 5.2, ¶ 1. Cho teaches that it is useful to continue gathering multiple copies of document collections because one of the copies may be unavailable later. In response to a user query, Cho discloses a "presentation filter" that "rolls up" collections so that "it only displays the link of one page in a collection, even if multiple pages within the collection satisfy the query." Thus, Cho keeps a record of duplicate documents which can be presented to a user, but only one is indexed and in the normal course only one is presented to the user.

The Examiner referred to §§ 5.1, 5.2 of Cho in the Office Action dated 08/20/2007, and continues to refer to § 5.1 of Cho in the Office Action dated 07/10/2008. These sections of Cho do not teach changing and indexing the representative for a set of duplicate documents. The Examiner's citation to "newly replicated collection, page 365, first column, second paragraph" is § 5.1 of Cho, which discloses only a replica avoiding process. The first copy is indexed, and remains the representative permanently. Thus, Cho does not teach having a set of duplicate documents where there is a representative document that changes and is indexed.

E. **The Examiner Agreed that Meyerzon and Cho do not Teach the Limitation Addressed Above**

In response to Applicants' arguments regarding Meyerzon and Cho, the Examiner cited the Wang reference, and indicated that it "solves the shortcomings described by the Applicant." Office Action dated 07/10/2008 at Page 19, ¶ 8. The "shortcomings" described by the Applicants' were that Meyerzon and Cho do not teach "indexing the representative document when the representative document is the newly crawled document; and repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed." The Applicants and the Examiner therefore agree that Meyerzon and Cho do not teach this limitation.

F. **Wang Does Not Teach This Important Limitation**

Wang provides a summary of "web search services." In section 3.1 Wang describes web crawlers, and in section 3.2 Wang describes how indexes are built to facilitate subsequent searching. Web crawlers detect new web pages as well as old web pages whose contents have changed. In either of these cases the web pages are indexed.

Wang's teaching to reindex a web page whose content has changed does not teach the limitations of the claims in the present application. It should go without saying that a web page whose content has changed is a not a duplicate of the earlier version of the same web page. In Wang there is no set of documents sharing a common document identifier; there is no representative for the set of documents; and there is no teaching to change the representative document. In Wang each document is distinct: it is periodically crawled to determine if it has changed, and reindexed if it has changed. There is no relationship between a document and any other document.

Moreover, the cited Meyerzon reference clarifies the key difference between reindexing a web page that changes versus detecting duplicate web pages to avoid indexing multiple copies of the same document. At column 2, lines 10-11 in the Background section, Meyerzon notes that "If a web page changes, then the index is updated with new information." But at column 2, lines 40-43 (still in the Background section), Meyerzon points out that "There is a need for an improved method and system for identifying duplicate documents, and using this information to avoid unnecessarily retrieving and processing such duplicates."

Thus both common sense and the Meyerzon reference dictate that ordinary indexing of changed web pages does not teach "indexing the representative document when the representative document is the newly crawled document; and repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed."

G. Combining Wang with Meyerzon and/or Cho Does Not Teach This Important Limitation

First, the Examiner has excised the phrase "such that at least some of the newly crawled documents are determined to be representative documents and are indexed" from the sentence where it appears. Taking the excised phrase in isolation, any web page without duplicates is trivially a "representative" of itself, and thus indexed when it changes. That is the situation described in Wang. But the claims in the present application explicitly exclude this trivial case:

> … newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed.

Claim 12, last paragraph. That is, the claim language applies only to newly crawled web pages for which there are duplicates. Wang's teaching regarding non-duplicate web pages does not address duplicate web pages, selecting representative web pages for sets of duplicates, or changing the representative document after an original representative document has been selected.

Second, a fair combination of Meyerzon, Cho, and Wang teaches nothing more that Meyerzon and Cho. Each newly crawled document is either a duplicate or a distinct non-duplicated web page. Wang addresses only the case of non-duplicated web pages, which is similarly addressed in Meyerzon at column 2, lines 10-11. That is, the non-duplicated web page is indexed if its content changes (and, of course, the new version is not a duplicate of any other web page). Duplicated web pages are addressed by Meyerzon, which applies the "first copy wins" rule. Wang's teaching for non-duplicated web pages thus can be combined

with Meyerzon and Cho; but Wang does not teach or suggest altering Meyerzon's "first copy wins" rule for handling duplicate web pages.

CONCLUSION

In light of the arguments presented above, Applicants respectfully request that the Examiner reconsider this application with a view towards allowance. The Examiner is encouraged to call the undersigned attorney at (650) 843-4000 should any issues remain unresolved.

Respectfully submitted,

Date: October 9, 2008 / Gary S. Williams / 31,066

Gary S. Williams
MORGAN, LEWIS & BOCKIUS LLP
2 Palo Alto Square
3000 El Camino Real, Suite 700
Palo Alto, California 94306
(650) 843-4000